UNIVERSIDADE FEDERAL DO RIO GRANDE DO SUL
INSTITUTO DE MATEMATICA
DEPARTAMENTO DE ESTATÍSTICA

IMPUTAÇÃO DE DADOS UTILIZANDO O ALGORITMO EM E REGRESSÃO LINEAR NO SPSS

Usiara Britto
Orientadora: Luciana Neves Nunes

Monografia apresentada para obtenção do grau de Bacharel em Estatística

Porto Alegre, dezembro de 2005

RESUMO

Neste trabalho é exposta a importância da imputação de dados. Para tanto, dois métodos de imputação foram estudados: Algoritmo EM e Método de Regressão Linear. Os métodos foram aplicados através de rotinas já implantadas no SPSS, no módulo "*Missing Value Analysis*". Como exemplo prático foram usados os dados da América Latina retirados do banco de dado "World95" contido no SPSS. Ainda, nesta monografia foi feita a imputação de dados através do método de Regressão Linear, feita pela rotina específica de regressão linear ao qual foi chamada de imputação por "regressão manual". Ao final do trabalho é apresentada uma comparação dos métodos, especificando suas semelhanças e diferenças nos resultados obtidos após a imputação.

AGRADECIMENTOS

*Aos meus **pais**, que estão sempre me mostrando a luz no fim do túnel,*

*aos meus **irmãos** que secam as minhas lagrimas e me fizeram voltar realidade,*

nas varias fugas que tive.

*Aos meus **amigos**, **colegas de trabalho** que continuaram acreditando em mim,*

mesmo quando, mesma não acreditava mais em mim.

*A minha orientadora, e amiga **Luciana** (Lu)*

*e ao **Lucas** que veio por último para somar forças. E a todos seres de luz que me*

acompanham nesta caminhada.

INDICE

1. INTRODUÇÃO 9

1.1. Objetivos 11

2. IMPUTAÇÃO DE DADOS 13

2.1 Classificação dos mecanismos de não-resposta 14

2.1.1. MCAR – *Missing Completely at Random* 15

2.1.2. MAR – *Missing at Random* 15

2.2. Métodos de imputação 16

2.2.1 Algoritmo EM - Expectation Maximization 16

2.2.2. Especificação do Algoritmo EM 18

2.2.3. Regressão Linear Múltipla 20

3. CORRELAÇÃO LINEAR 22

4. ANÁLISE DE DADOS AUSENTES PELO SPSS VERSÃO 11.5 24

5. BANCO DE DADOS 29

5.1. Correlação 38

6. EXEMPLO 40

6.1. Imputação por Algoritmo EM 42

6.2. Imputação por Regressão Linear 45

6. 3. Regressão Linear – Manualmente 48

7. CONCLUSÃO 53

8. CONSIDERAÇÕES 55

9. BIBLIOGRAFIA _____ **57**

10. ANEXOS _____ **60**

ÍNDICE DE TABELAS

Tabela 1: Interpretação da correlação ____ 23

Tabela 2: Estatísticas Descritivas Mundiais (n=109) ____ 30

Tabela 3: Estatísticas Descritivas da África (n=19) ____ 31

Tabela 4: Estatísticas Descritivas da América Latina (n=21) ____ 32

Tabela 5: Estatísticas Descritivas da Ásia (n= 19) ____ 33

Tabela 6: Estatísticas Descritivas da Europa/EUA/Canadá (n=19) ____ 35

Tabela 7: Estatísticas Descritivas da Europa Oriental (n = 14) ____ 36

Tabela 8: Estatísticas Descritivas do Oriente Médio (n= 17) ____ 37

Tabela 9: Resumo da Matriz de Correlação ____ 39

Tabela 10: Variáveis correlacionadas com mortalidade infantil ____ 43

Tabela 11: Valores imputados (em negrito) pelo algoritmo EM para mortalidade infantil (Correlação Negativa) ____ 43

Tabela 12: Valores imputados (em negrito) pelo algoritmo EM para mortalidade infantil (Correlação Positiva) ____ 44

Tabela 13: Valores imputados (em negrito) pela regressão linear para mortalidade infantil (Correlação Negativa) ____ 47

Tabela 14:Valores imputados (em negrito) pela regressão linear para mortalidade infantil (Correlação Positiva) ____ 47

Tabela 15: Valores imputados (em negrito) pela regressão linear manual para mortalidade infantil (Correlação Negativa) 51

Tabela 16: Valores imputados (em negrito) pela regressão linear manual para mortalidade infantil (Correlação Positiva) 52

Anexo 1: Matriz de Correlação – Geral 61

Anexo 2: Matriz de Correlação – América Latina 64

ÍNDICE DE FIGURAS

Figura 1: Análise dos Dados ____ **40**

Figura 2: *Missing Value Analysis* ____ **41**

Figura 3: *Missing Value Analysis* – Variáveis para EM e Regressão ____ **41**

Figura 4: *Missing Value Analysis*: Algoritmo EM ____ **42**

Figura 5: Análise ____ **45**

Figura 6: Definição das Variáveis ____ **45**

Figura 7: Definição das variáveis (Dependentes e Independentes) ____ **46**

Figura 8: Definição do Ajuste ____ **46**

Figura 9: Análise ____ **48**

Figura 10: Regressão Linear ____ **49**

Figura 11: Regressão Linear - Estatísticas ____ **49**

Figura 12: Regressão Linear - Gráficos ____ **50**

Figura 13: Regressão Linear: Salvar ____ **50**

Figura 14: Regressão Linear ____ **51**

1. INTRODUÇÃO

É muito comum ao se coletar dados amostrais que algumas das informações fiquem ausentes. Isto ocorre quando os dados foram respondidos, mas percebe-se claramente que estas informações estão erradas e por este motivo precisam ser descartadas, ou não foram mesmo respondidos. Em muitas situações é impossível entrar em contato com o respondente a fim de corrigir estes dados. Esta é uma das motivações para a realização desse trabalho que trata da imputação de dados, ou seja, substituição dos dados ausentes, pois a falta dessas informações pode trazer perdas enormes na análise dos dados e pode comprometer a pesquisa. Através de uma revisão da literatura foram investigados métodos de imputação utilizados para solucionar este problema.

Segundo Mattson (2005), o tratamento de dados ausentes requer uma compreensão das características dos mesmos e o provável impacto que estaria ocasionando na análise. Se após a realização de uma imputação perceber-se que o impacto no conjunto de dados é grande, pode-se concluir que é mais apropriado ignorar os dados ausentes e fazer uma análise desconsiderando-os. A imputação não pode distorcer os dados iniciais, nem mesmo aumentar a discrepância ou distorcer as estimativas, ou seja, as estimativas obtidas com tais análises podem ser viesadas, especialmente se os indivíduos que são incluídos na análise são sistematicamente diferentes daqueles que foram excluídos em termos de uma ou mais variáveis. Para contornar esse problema, desde os anos 80 surgiram técnicas estatísticas que envolvem

imputação de dados ausentes. Essas técnicas têm por objetivo "completar" os bancos de dados e possibilitar a análise com todos os indivíduos do estudo.

Os primeiros métodos de imputação de dados eram relativamente simples, pois utilizavam técnicas que não necessitavam de grandes recursos computacionais, por exemplo, substituição pela média, mediana, moda, ou seja, substituir o dado ausente pela média dos dados presentes. Com o avanço tecnológico foram criadas condições para que métodos mais sofisticados pudessem ser utilizados para a imputação de dados. Métodos de máxima verossimilhança, tais como algoritmo EM (*Expectation Maximization*) e regressão linear passaram a ser usados com esse objetivo. O presente trabalho apresentará esses dois métodos.

A regressão linear pode ser simples ou múltipla, dependendo do número de variáveis independentes utilizadas no modelo. Quando há uma variável independente a regressão é simples, caso tenha duas ou mais variáveis independentes a regressão é considerada múltipla. Caso existam dados ausentes, é preciso definir quais variáveis serão utilizadas como dependentes ou independentes. Para isso, verifica-se entre todas as variáveis quais são as mais correlacionadas. Das mais correlacionadas, a variável que possuir o maior número de valores ausentes será a variável dependente e as demais serão consideradas variáveis independentes. O algoritmo EM é um processo computacional iterativo para obtenção de estimativas de máxima verossimilhança, especialmente útil em situações de dados incompletos. A escolha das variáveis que serão utilizadas no algoritmo EM é feita analogamente ao método de regressão linear.

Existem, ainda, os métodos de imputação múltipla que foram desenvolvidos mais recentemente. Esses métodos surgiram para resolver situações em que a imputação

simples não era suficiente. Porém, nesse trabalho esse enfoque não será abordado. Rubin (1987)

Em pesquisas onde são usados bancos de dados públicos, ou seja, dados que são compartilhados por muitos usuários, a imputação pode ser uma boa solução para o problema de dados ausentes, pois não há a possibilidade de se recuperar esta informação. O banco de dados utilizado nesse trabalho é proveniente do SPSS. No SPSS existem vários bancos de dados públicos disponíveis, dentre eles, o World95, que foi escolhido por conter variáveis demográficas que traçam o perfil de 109 países. Para esse trabalho foram selecionados os dados referentes a América Latina, levando-se em conta a semelhança dos países que a compõe quanto à sua colonização, para a realização da imputação de dados. Entretanto, o banco completo, ou seja, todas as regiões mundiais foram descritas.

Após a realização das imputações dos dados pelos diferentes métodos, serão feitas comparações, através das estimativas da média e do desvio padrão. Tais comparações têm por objetivo avaliar a qualidade das imputações.

1.1. Objetivos

Objetivo Geral:

- ✓ Descrever e comparar métodos de imputação implantados no SPSS 11.5, através de um banco de dados do próprio SPSS (World95).

Os objetivos específicos:

- ✓ Fazer a imputação por Algoritmo EM utilizando a rotina *"Missing value Analysis"* do SPSS;
- ✓ Fazer a imputação por Regressão Linear utilizando a rotina *"Missing value Analysis"* do SPSS;
- ✓ Fazer a imputação por Regressão Linear utilizando a rotina de regressão linear do SPSS;
- ✓ Comparar os resultados das imputações.

2. IMPUTAÇÃO DE DADOS

O problema da ausência parcial ou total de alguma informação deixa o pesquisador com dúvida sobre qual a melhor solução para este problema: Se elimina todas as demais informações desta variável ou decide por um método de imputação de dados ausentes. Este problema já vem de longos anos, em 1930, Allan e Wishart, que trabalhavam com pesquisas agronômicas, introduziram fórmulas para imputar valores em uma única observação ausente em planejamento de blocos aleatorizados e quadrados latinos. Mais tarde, em 1933 Yates estudou o caso da falta de múltiplas observações e mostrou que o valor ausente pode ser obtido através da minimização da soma de quadrados dos resíduos. Da agricultura, a imputação dos dados estendeu-se para outras áreas, pois este problema não é um privilégio somente das pesquisas agronômicas, mas sim de todas as áreas de pesquisa (DORNELES, 1999).

Outro problema é saber qual a proporção ideal de dados ausentes para se ter uma boa imputação dos dados. Se a proporção de valores ausentes for pequena, por exemplo, até 5%, então uma única imputação é uma boa opção. Se essa proporção estiver entre 5% e 15%, a imputação única pode ser bastante razoável , porém a imputação múltipla também pode ser utilizada. Para uma proporção maior do que 15% a imputação múltipla é preferível (DORNELES, 1999).

A conclusão de uma única imputação tende a exagerar a precisão, porque subestima a variabilidade do conjunto de dados, implicando assim em um decréscimo na variabilidade da variável imputada. Há vários métodos de imputação de dados, sendo os mais conhecidos e utilizados os da Média Geral, onde simplesmente é atribuída a

média geral das informações presentes para todos os valores ausentes desta variável; o Aleatório Geral, onde é atribuído ao valor ausente uma informação selecionada aleatoriamente da amostra total das informações; o da Média dentro das classes, este método considera a amostra total dividida em classes de imputação de acordo com valores das variáveis auxiliares e é atribuído ao valor ausente a média da classe. Na verdade há uma gama de métodos de imputação, porém para se decidir sobre qual usar deve-se ter em mente que a imputação não deve reduzir a variabilidade da variável imputada, pois esta variabilidade reduzida pode influenciar as estimativas dos efeitos das variáveis.

Para se decidir sobre qual método de imputação utilizar, deve-se verificar qual o mecanismo de não-resposta dos dados ausentes.

2.1 Classificação dos mecanismos de não-resposta

Segundo Silva (1989), a maioria das técnicas disponíveis de imputação assumem que os valores ausentes são "perdidos ao acaso". Esta suposição é matematicamente conveniente porque permite evitar um modelo de probabilidade explícito para ausência deste dado. Em algumas aplicações, porém, ignorar esta suposição pode acarretar sérios problemas. Com base em um estudo longitudinal, por exemplo, é possível que os dados ausentes estejam relacionados a valores de dados atuais. Imputações podem ser criadas com os princípios abaixo, definições 2.1.1. e 2.1.2, ou seja, faz-se uma suposição ou assume-se um modelo para o mecanismo de dados ausentes, e as conclusões resultantes serão validadas após a escolha do mecanismo. O método de imputação é classificado

pelo tipo de variável, o mecanismo pode contar com uma ou mais variáveis auxiliares ou somente a variável em questão.

2.1.1. MCAR – *Missing Completely at Random*

Quando o mecanismo de não-resposta não depende da variável de interesse, diz-se que os dados são ausentes por um processo totalmente aleatório, ou seja, os dados ausentes que não tem nenhuma "ligação" com outra variável presente.

Exemplo: Medidas de laboratório que foram perdidas por causa de um tubo de teste derrubado. Em estudo posterior não haverá a resposta dessa variável, entretanto esta não-resposta não tem nenhuma conexão com alguma das variáveis presentes.

2.1.2. MAR – *Missing at Random*

Quando o mecanismo de não-resposta depende de uma ou mais informação de outras variáveis presentes, ou seja, a probabilidade dos dados perdidos em uma particular variável depende de outras variáveis disponíveis.

Exemplo: Considere uma pesquisa na qual as mulheres são menos prováveis de informar sua renda pessoal. Se soubermos o sexo de todos os respondentes e tivermos a renda para algumas mulheres, podem ser feitas estimativas de renda para aquelas que não responderam através da variável sexo.

2.2. Métodos de imputação

A imputação pode ser feita sem a utilização de variáveis auxiliares, ou seja, a imputação dos valores ausentes para uma determinada variável é baseada somente nos dados observados. Quando a imputação utiliza variáveis auxiliares, a previsão dos valores ausentes utiliza outras variáveis do mesmo estudo.

Dos diversos métodos de imputação, nesse trabalho serão detalhados os métodos de Algoritmo EM e o da Regressão Linear.

2.2.1 Algoritmo EM - Expectation Maximization

Segundo Luna (2004), o algoritmo EM surgiu da unificação de uma série de trabalhos aparentemente sem relação e apresentado com esse nome por Demspter *et al* (1977). Depois disto, este algoritmo permaneceu por muito tempo esquecido. Basicamente, se uma variável foi algumas vezes observada e outras não, pode-se utilizar os casos para os quais foi observada para "aprender" a predizer seus valores quando não foram observados. O algoritmo EM realiza esta tarefa, sempre que seja conhecida a forma geral da distribuição de probabilidade das variáveis.

Em geral, os parâmetros descrevem as características de uma população. Seus valores são estimados de amostras coletadas dessa população. O algoritmo EM faz uma estimativa de máxima verossimilhança dos parâmetros, ou seja, estima parâmetros que sejam os mais consistentes com os dados da amostra no sentido de maximizar a função de verossimilhança.

Antes de introduzir o algoritmo EM, será definida a função de verossimilhança e o estimador de máxima verossimilhança, conforme Mood e Boes (1974), para o caso geral de n variáveis aleatórias que dependem do parâmetro θ, que pode ser um vetor de parâmetros.

Definição 1: Função de Verossimilhança

A função de verossimilhança de n variáveis aleatórias $X_1, X_2, \ldots, X_n$ *é definida como a densidade conjunta de n variáveis aleatórias, digamos* $fx_1,\ldots,x_n\ (x_1, \ldots, x_n; \theta)$, *considerada como função de* θ. *Em particular, se* $X_1, \ldots, X_n$ *é uma amostra aleatória da densidade* $f(x; \theta)$, *então a função de verossimilhança é* $\prod_{i=1}^{n} f(x_i;\theta)$.

Definição 2: Estimador de Máxima Verossimilhança

A função de verossimilhança para as variáveis aleatórias $X_1, X_2, \ldots, X_n$. *Se* θ, *onde* $\hat{\theta} = (x_1, x_2, \ldots, x_n)$ *é uma função das observações* $[x_1, \ldots, x_n]$ *é o valor de* θ *em* $\hat{\theta}$ *(universo dos* θ*'s) que maximiza* $L(\theta)$, *então* $\hat{\theta} = \theta(X_1, X_2, \ldots, X_n)$ *é o estimador de máxima verossimilhança de* θ. *O parâmetro é estimado pela estimativa de máxima verossimilhança.*

2.2.2. Especificação do Algoritmo EM

O algoritmo EM pode ser aplicado nos casos onde se deseja estimar algum conjunto de parâmetros que descreve uma certa distribuição de probabilidades conjunta, dada somente uma parte observada dos dados produzidos por esta distribuição. Para ser mais preciso, considere que $X = \{x_1, \ldots, x_m\}$ denote os dados observados em um conjunto de m instâncias ocorridas independentemente, seja $Z = \{z_1, \ldots, z_m\}$ os dados não observados nestas mesmas instâncias, e seja $Y = (X \cup Z)$ o total de dados. Z pode ser tratada como uma variável aleatória cuja distribuição de probabilidades depende do conjunto de parâmetros desconhecido θ e dos dados observados X. Analogamente, Y é uma variável aleatória porque é definida em termos da variável aleatória Z (MITCHELL, 1997).

Para descrever a forma geral do algoritmo EM, h denota a hipótese atual dos parâmetros θ e h' denota a hipótese revisada que é estimada em cada iteração do algoritmo EM. O algoritmo EM busca a hipótese h' de máxima verossimilhança, isto é, que maximize $E[\ln P(Y|h')]$. Este valor esperado é calculado sob a distribuição de probabilidade de Y.

Mas, o que significa $E[\ln P(Y|h')]$?

- Primeiro, $P(Y|h')$ é a verossimilhança de todos os dados Y, dada a hipótese h'. É razoável que se queira encontrar h' que maximize alguma função desta quantidade;
- Segundo, maximizar o logaritmo desta quantidade $\ln P(Y|h')$ também maximiza $P(Y|h')$;

- Terceiro, se introduz o valor esperado de $E[\ln P(Y|h')]$ porque o total de dados Y *é*, ele próprio, uma variável aleatória.

Dado que os dados Y são uma combinação dos dados observados X e não observados Z, deve-se mediar sobre os possíveis valores não observados Z, ponderando cada um de acordo com suas probabilidades. Em outras palavras, toma-se o valor esperado $E[\ln P(Y|h')]$ sobre a distribuição de probabilidade da variável aleatória Y. A distribuição de probabilidades de Y é determinada pelos valores completamente conhecidos para X, mais a distribuição de probabilidade de Z.

Qual é a distribuição de probabilidade de Y? Em geral, não se conhece esta distribuição porque ela é determinada pelos parâmetros θ que se quer estimar. Entretanto, o algoritmo EM usa sua hipótese atual h no lugar do parâmetro θ atual para estimar a distribuição de probabilidades de Y. Considere a definição de uma função $Q(h'|h)$ que dá $E[\ln P(Y|h')]$ como uma função de h', sob a suposição que $\theta = h$ e dada a porção observada X dos dados Y.

$$Q(h'|h) = E[\ln P(Y|h') \mid h, X]$$

Escreve-se esta função Q na forma $Q(h'|h)$ para indicar que ela é definida em parte pela suposição que a hipótese atual h *é* igual a θ.

Em sua forma geral, o algoritmo EM repete os dois passos seguintes, até a convergência:

> **Passo E**: *Expectation(E):* calcula $Q(h'|h)$ usando a hipótese atual h e os dados observados X para estimar a distribuição de probabilidade sobre Y.
> $Q(h'|h) = E[\ln P(Y|h') \mid h, X]$

Passo M: *Maximization(M):* troca a hipótese h pela hipótese h' que maximiza esta função ao Q. $h = \arg\max Q(h' \mid h)\, h'$

Quando a função Q é contínua, o algoritmo EM converge para um ponto estacionário da função de verossimilhança $P(Y|h')$. Quando esta função tem um único máximo, EM convergirá para esta estimativa de máxima verossimilhança global para h'. De outro modo, o algoritmo converge somente para um máximo local.

2.2.3. Regressão Linear Múltipla

O método de imputação que utiliza regressão linear tem a pressuposição implícita de que a variável que tem a ocorrência de valores ausentes pode ser considerada como uma variável dependente das demais. Quando se considera uma variável dependente e k variáveis explicativas ou independentes, a equação de regressão é dada por:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_k X_k + \varepsilon$$

Sendo β_0, β_1,..., β_k os coeficientes de regressão, estimados a partir dos dados não ausentes, e suas estimativas (b_0,b_1,...,b_k), utilizadas para prever um valor ausente.

Quando se realiza imputação utilizando-se regressão linear múltipla, a previsão do i-ésimo valor ausente é escrita como:

$$\hat{Y} = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_k x_k$$

podendo a imputação ser feita de modo determinístico quando o resíduo ε assume sempre valor zero (método da imputação da média regressão), ou de modo estocástico, quando o resíduo ε é sorteado de um banco formado pelos resíduos brutos do modelo ajustado, para os respondentes (método da imputação da média da regressão mais efeito de aleatoriedade).

3. CORRELAÇÃO LINEAR

Quando existirem duas séries de dados, existirão várias medidas estatísticas que podem ser usadas para capturar como as duas séries se movem juntas através do tempo. As duas mais largamente usadas são a correlação e a covariância. Para duas séries de dados, X (X_1, X_2,..) e Y(Y_1,Y_2...), a covariância fornece uma medida não padronizada do grau no qual elas se movem juntas, e é estimada tomando o produto dos desvios da média para cada variável em cada período.

$$Cov(X,Y) = E(XY) - E(X)*E(Y)$$

O sinal na covariância indica o tipo de relação que as duas variáveis tem. Um sinal positivo indica que elas se movem juntas e um negativo que elas se movem em direções opostas. Enquanto a covariância cresce com o poder do relacionamento, ainda é relativamente difícil fazer julgamentos sobre o poder do relacionamento entre as duas variáveis observando a covariância, pois ela não é padronizada.

A correlação é a medida padronizada da relação entre duas variáveis. Ela pode ser definida da covariância:

$$r = \frac{Sxy}{SxSy}$$

onde

$$S_{xy} = \frac{\Sigma(x_i - \bar{x})(y_i - \bar{y})}{n - 1}$$

A correlação nunca pode ser maior do que 1 ou menor do que menos 1. Uma correlação próxima a zero indica que as duas variáveis não estão relacionadas linearmente. Uma correlação positiva indica que as duas variáveis movem-se juntas, e a relação é forte quanto mais a correlação se aproxima de um. Uma correlação negativa indica que as duas variáveis movem-se em direções opostas, tendo uma relação mais forte quando próxima se de – 1.

A Tabela 1 descreve uma forma de como pode-se interpretar a correlação É claro que as interpretações dependem de cada contexto em particular.

Tabela 1: Interpretação da correlação

VALOR P (+ OU -)	INTERPRETAÇÃO
0,00 a 0,19	Uma correlação bem fraca
0,20 a 0,39	Uma correlação fraca
0,40 a 0,69	Uma correlação moderada
0,70 a 0,89	Uma correlação forte
0,90 a 1,00	Uma correlação muito forte

4. ANÁLISE DE DADOS AUSENTES PELO SPSS VERSÃO 11.5

O pacote estatístico SPSS. tem uma rotina que realiza a imputação de dados ausentes. Essa rotina é denominada *Missing Value Analysis*. Neste estudo serão mostrados os métodos de imputação disponíveis que são a regressão linear e o algoritmo EM.

Quando há valores ausentes no banco de dados, a dificuldade em se obter estimativas válidas do parâmetro depende do mecanismo de não-resposta. Como mencionado nas seções 2.1.1 e 2.1.2, existem os mecanismos MCAR e MAR. Quando a ausência depende dos valores observados e ausentes, tem-se o mecanismo NMAR (*Not Missing at Random*). O analista pode ignorar o mecanismo se os dados são provenientes do MCAR ou MAR, porém se os dados forem NMAR o mecanismo não poderá ser ignorado. Na prática, entretanto, pode ser bastante difícil identificar o real mecanismo de não-resposta.

Se os valores ausentes são MCAR, as estimativas obtidas pelo SPSS não serão viesadas. Mas isto não é impressionante, porque os valores observados são uma amostra aleatória da população. Se os valores ausentes forem NMAR, as estimativas podem ser viesadas. Isto não é surpreendente, pois o *Missing Value Analysis* não incorpora o mecanismo de não-resposta. Quando os valores ausentes são MAR, é possível se obter estimativas não viesadas, no entanto a maioria das estimativas obtidas pelo *Missing Value Analysis* são viesadas (HIPPEL, 2004).

4.1. Teste de Little´s MCAR

O teste de Little, desenvolvido em 1988, está implementado no *Missing Value Analysis*. Este teste tem como hipótese nula que os dados ausentes são MCAR. Por exemplo, se o pesquisador suspeita que os dados ausentes foram perdidos através do mecanismo de MAR ou MCAR, pode realizar o teste e a hipótese é rejeitada ($p < 0,05$), terá mais segurança de que os dados são MAR.

1

4.2 Métodos de eliminação do *Missing Value Analysis*

Embora haja problemas na definição do mecanismo de não-resposta, muitas vezes os pesquisadores desejam estimar os parâmetros quando os valores ausentes foram perdidos ao acaso. O *Missing Value Analysis* oferece quatro métodos para este propósito. Os dois primeiros métodos estão baseados em desconsiderar os dados ausentes. Enquanto os outros dois métodos, algoritmo EM e regressão linear, fazem a imputação dos dados ausentes.

Os dois métodos que são de eliminação de dados, *listwise* e *pairwise*, já estavam implementados no SPSS básico. Estes métodos não são de imputação. Uma variedade de análises do SPSS, incluindo análises de regressão, correlação, fatorial e outras usam esses métodos. No *Missing Value Analysis*, esses métodos servem para fornecer as estimativas de ponto para médias, desvio padrão, e covariâncias.

4.2.1 Eliminação por *Listwise*

O método de eliminação *listwise* mostra as médias, a matriz de correlação e a matriz de covariâncias, omitindo os casos que têm valor ausente em qualquer uma das variáveis consideradas.

4.2.2 Eliminação por *Pairwise*

O método de eliminação *pairwise* mostra para cada par de variáveis quantitativas o número de pares de valores observados, e a média, variância, covariância e correlação dos pares. Cada cálculo é realizado usando todos os valores para os quais ambas variáveis tem valores presentes.

4.3. Métodos de imputação por verossimilhança

Os métodos apresentados acima, de eliminação, não são considerados métodos de imputação de dados, tanto que estes, já estavam disponíveis em software SPSS Básico. A complementação, propriamente dita, ocorreu nos métodos baseados em probabilidade. Embora a imputação e métodos estejam baseados na Verossimilhança eles têm um potencial para produzir estimativas, cujo os erros padrão são próximos do verdadeiro valor. Quando os valores são MAR, o único método que produz resultados válidos no *Missing Value Analysis* é o Algoritmo EM. A implementação de *Missing Value Analysis* deste método é limitado para as estimativas de médias, desvio padrão e covariações.

4.3.1 Imputação de regressão linear

Na imputação pela regressão (IR), um modelo de regressão é usado para imputar (ou preencher) os valores ausentes. Para que ocorra a imputação são usados métodos convencionais para derivar estimativas dos dados completos e imputar os dados ausentes.

A IR tem algumas dificuldades práticas. Um problema é que podem haver muitos regressores possíveis. A segunda dificuldade é que os regressores podem possuir muitos valores ausentes. A identificação destes problemas pode ser feita através de diferentes equações de regressão.

Aplicado aos dados, a IR do *Missing Value Analysis* realiza a imputação através do ajuste do modelo abaixo:

$$Y_{imp} = a + bX + \varepsilon \text{ onde } \varepsilon \sim N(0, \sigma^2)$$

O modelo mostrado apresenta resíduos normais. Mas o *Missing Value Analysis* também pode imputar valores que usam resíduos de t, resíduos de reamostragem, ou não utilizar nenhum resíduo. Análises posteriores podem ser feitas com o banco de dados completo pelas imputações.

4.3.2 Imputação pelo algoritmo EM

O segundo método de imputação disponível no *Missing Value Analysis* é o Algoritmo EM (Dempster, Laird, e Rubin 1977). O Algoritmo EM está baseado em iterar o processo de imputação de regressão, a fim de ter as primeiras estimativas a

serem utilizadas num segundo momento. Hippel (2004) mostra um exemplo, onde o primeiro passo do EM usa a seguinte regressão inicial:

$$Y_{imp} = a_0 + b_0X$$

O *Missing Value Analysis* usa as estimativas de *pairwise* para "a" e "b". Note que as imputações não incluem nenhuma variação residual. Os resíduos acrescentariam distorções ao algoritmo, introduzindo variação aleatória que não seriam inerentes aos dados.

Usando os dados observados e imputados, o método EM re-estima as médias, desvios padrão e covariâncias, usando uma fórmula que compensa a falta de variação residual nos valores imputados de Y. Estas novas estimativas de regressão são usadas para gerar imputações novas de Y, e o processo itera até a convergência. O *Missing Value Analysis* fixa um limite de 25 iterações, mas gera uma mensagem de erro caso o limite não seja suficiente.

Embora assintoticamente não viesado, as estimativas do algoritmo EM são de uso limitado. Na maioria das análises, os pesquisadores não têm interesse em médias, desvio padrão, ou covariâncias, mas nos parâmetros de algum modelo causal ou descritivo. Às vezes estes parâmetros podem ser calculados transformando as médias, desvios padrão ou covariâncias fornecidas pelo algoritmo EM. Entretanto, seria conveniente para analisar os dados imputados o uso do próprio SPSS ou outro software.

5. BANCO DE DADOS

Segundo Hauser e Duncan demografia é um estudo das mudanças e dos componentes de tais mudanças da população, por isto baseia-se no tamanho, na distribuição geográfica e na composição da população. Essas mudanças podem ser identificadas como natalidade, mortalidade, movimentos territoriais (migração) e mobilidade social (mudança de status). Suas informações são utilizadas para definir características de uma população e através da análise destes dados são estabelecidas políticas governamentais em educação, habitação, saúde e economia, entre outras.

Para este trabalho, foi usado como exemplo de imputação o banco de dados World95, integrante do SPSS. Este banco de dados contém uma descrição das variáveis que medem o grau de desenvolvimento de um país. As variáveis contidas são: localidade, população em milhões, densidade populacional, população urbana, religião predominante do país, expectativa de vida masculina e feminina, índice de alfabetização, taxa de crescimento populacional, mortalidade infantil, PIB, classificação do país por região ou grupo econômico, quantidade calórica consumida, número de casos de Aids, natalidade, mortalidade, incidência de Aids, taxa de natalidade por mortalidade, fertilidade, índice de alfabetização masculino e feminino, clima da localidade e região geográfica. Nas tabelas a seguir, estão as estatísticas descritivas das variáveis.

A partir da tabela 2 identifica-se, através das médias e os desvios padrão das variáveis em estudo, o perfil mundial, incluindo 109 países. Observa-se a média de 47.723 mil habitantes para a população mundial, expectativa média de vida de 70 anos

para a população feminina e 65 para a população masculina, vê-se ainda que em média, 56,53% da população mundial reside na região urbana e 78,34% são alfabetizadas. Aparentemente, estes dados retratam uma realidade mascarada e para verificar esse fato fez-se uma comparação dos resultados mundiais com os resultados de regiões demográficas definidas.

Tabela 2: Estatísticas Descritivas Mundiais (n=109)

Variáveis	Média	Desvio Padrão
População (mil)	47.723,88	146.726,36
Densidade Populacional (hab./Km)	203,41	675,71
População Urbana(%)	56,53	24,20
Expectativa de vida feminina (anos)	70,16	10,57
Expectativa de vida masculina (anos)	64,92	9,27
Alfabetização (%)	78,34	22,88
Crescimento Populacional (%)	1,68	1,20
Mortalidade Infantil (%)	42,31	38,08
Produto Interno Bruto ($)	5.859,98	6.479,84
Calorias (Cal)	2.753,83	567,83
Aids (nº de casos)	7.914,26	40.690,67
Natalidade (nat/1000)	25,92	12,36
Mortalidade (mort/1000)	9,56	4,25
Índice de Aids (Aids/100000)	24,38	49,45
Taxa de natalidade por mortalidade	3,20	2,12
Fertilidade (%)	3,56	1,90
Alfabetização Masculino (%)	78,73	20,45
Alfabetização Feminino(%)	67,26	28,61

A primeira região demográfica analisada foi o Continente Africano (tabela 3) onde vê-se que 29,26% da população mora no meio urbano, apenas 47,26% de sua

população é alfabetizada, as pessoas do sexo feminino têm expectativa de vida de 54,26 anos e os homens com de 47,26 anos e tem-se ainda um indicador péssimo para a Aids onde 75,75% da população encontra-se contaminada com este vírus. É triste revelar que a população residente neste continente tem um nível de qualidade de vida nada agradável, estando muito abaixo da qualidade de vida mundial. A África apresenta um índice assustador quanto à mortalidade infantil, pois este continente tem 94,18% de suas crianças mortas antes de completar um ano de vida.

Tabela 3: Estatísticas Descritivas da África (n=19)

Variáveis	Média	Desvio Padrão
População (mil)	19.757,11	24.357,86
Densidade Populacional (hab./Km)	62,19	77,85
População Urbana(%)	29,26	15,06
Expectativa de vida feminina (anos)	54,26	7,98
Expectaticva de vida masculina (anos)	50,79	7,26
Alfabetização (%)	47,26	17,86
Crescimento Populacional (%)	2,72	0,46
Mortalidade Infantil (%)	94,18	28,65
Produto Interno Bruto ($)	998,68	1.178,26
Calorias (Cal)	2.152,31	208,23
Aids (nº de casos)	10.114,26	14.264,73
Natalidade (Nat/1000)	42,00	6,41
Mortalidade (mort/1000)	14,74	5,05
Índice de Aids (Aids/100.000)	75,75	87,42
Taxa de natalidade por mortalidade	3,10	0,84
Fertilidade (%)	6,08	1,13
Alfabetização Masculino (%)	54,17	17,14
Alfabetização Feminino(%)	32,06	16,01

Analisando a América Latina (tabela 4) verifica-se que ela apresenta 61,38% de sua população sendo urbana, com expectativa de vida para as mulheres de aproximadamente 72 anos e para os homens 66 anos, tendo 82,67% de sua população alfabetizada e uma taxa de 23,20% de sua população contaminada pelo vírus da Aids. Infelizmente a América Latina ainda está longe de exterminar com a mortalidade infantil, pois tem 39,11% de taxa de mortalidade infantil.

Tabela 4: Estatísticas Descritivas da América Latina (n=21)

variáveis	**Média**	**Desvio Padrão**
População (mil)	21.928,86	36.806,78
Densidade Populacional (hab/km)	88,28	137,09
População Urbana(%)	61,38	17,96
Expectativa de vida feminina (anos)	71,76	7,39
Expectativa de vida masculina (anos)	66,24	7,33
Alfabetização (%)	82,67	13,48
Crescimento Populacional (%)	1,88	0,68
Mortalidade Infantil (%)	39,11	24,52
Produto Interno Bruto ($)	1.997,67	1.482,12
Calorias (Cal)	2.500,16	322,56
Aids (nº de casos)	4.589,43	11.002,19
Natalidade (Nat/1000)	26,90	6,80
Mortalidade (Mort/1000)	7,33	3,15
Índice de Aids (Aids/100.000)	23,20	32,59
Taxa de natalidade por mortalidade	4,10	1,57
Fertilidade (%)	3,34	1,06
Alfabetização Masculino (%)	85,10	12,27
Alfabetização Feminino (%)	80,62	15,15

A Ásia (tabela 5) apresenta em seus indicadores demográficos uma população dividida entre o campo e a cidade, ou seja, 48,95% da população reside na zona urbana. Quanto a expectativa de vida, a feminina é de aproximadamente 69 anos e para os homens é de 65 anos. Ainda, nesta população, o índice de Aids é de 3,01%, ou seja, enquanto existe um número elevado de casos de aids na África e na América Latina, no Continente Asiático este problema parece estar sob controle. Na Ásia, apesar de ter um índice baixo de casos de aids, há 49,06% de mortalidade infantil, mostrando a precariedade com a saúde de um cidadão que acaba de vir ao mundo.

Tabela 5: Estatísticas Descritivas da Ásia (n= 19)

Variável	Média	Desvio Padrão
População (mil)	61.393,05	324.404,84
Densidade Populacional	718,81	1.523,69
População Urbana(%)	48,95	29,70
Expectativa de vida feminina (anos)	68,74	11,00
Expectativa de vida masculina (anos)	64,53	8,82
Alfabetização (%)	75,74	25,16
Crescimento Populacional (%)	1,58	0,85
Mortalidade Infantil (%)	49,06	46,10
Produto Interno Bruto ($)	5.457,89	6.939,68
Calorias (Cal)	2.633,46	448,27
Aids (nº de casos)	717,22	1.650,63
Natalidade (nat/1.000)	25,14	11,64
Mortalidade(mort/1.000)	8,75	4,22
Índice de Aids (Aids/100.000)	3,01	6,83
Taxa de natalidade por mortalidade	3,04	1,02
Fertilidade (%)	3,23	1,75
Alfabetização Masculino (%)	79,19	21,23
Alfabetização Feminino (%)	64,56	31,01

A região chamada de Região Europa/EUA/Canadá.no banco de dados é composta pelos seguintes países: Canadá, EUA, Áustria, Bélgica, Dinamarca, Finlândia, França, Alemanha, Grécia, Islândia, Irlanda, Itália, Países Baixos, Noruega, Portugal, Espanha, Suécia, Suíça e Reino Unido. Seus resultados (tabela 6) vêm na contramão das três regiões apresentadas acima, pois 73,68% da população mora na cidade, apenas 6,78% de suas crianças morrem antes de completarem um ano, suas expectativas de vida, tanto feminina (80 anos) como masculina (74 anos), estão acima da expectativa de vida mundial, porém seu índice de crescimento populacional é menor do que as demais regiões, mostrando assim um envelhecimento de sua população. Porém, a Aids assombra os europeus como assombra os latinos americanos e os africanos, pois 30,13% de sua população está contaminada com este vírus. Portanto, pode-se resumir essa região como um continente que apresenta características boas como uma população quase que totalmente alfabetizada, pouca mortalidade infantil e uma expectativa de vida superior as demais, porém com um crescimento populacional muito baixo.

Tabela 6: Estatísticas Descritivas da Europa/EUA/Canadá (n=19)

Variável	Média	Desvio Padrão
População (mil)	35.445,42	59.677,42
Densidade Populacional (hab/km)	118,54	108,19
População Urbana(%)	73,68	15,32
Expectativa de vida feminina (anos)	80,11	1,24
Expectativa de vida masculina (anos)	73,74	1,19
Alfabetização (%)	97,47	3,49
Crescimento Populacional (%)	0,46	0,29
Mortalidade Infantil (%)	6,78	1,13
Produto Interno Bruto ($)	16.715,74	3.890,30
Calorias (Cal)	3.454,00	236,14
Aids (Nº de casos)	28.061,16	93.391,95
Natalidade (nat/ 1.000)	12,68	1,49
Mortalidade (mort/1.000)	9,81	1,22
Índice de Aids (Aids/100.000)	30,13	35,21
Taxa de natalidade por mortalidade	1,33	0,32
Fertilidade (%)	1,72	0,25
Alfabetização Masculino (%)	95,80	3,83
Alfabetização Feminino (%)	91,40	6,11

A Europa Oriental (tabela 7), que é composta por: Belarus, Bósnia, Bulgária, Croácia, Republica Tcheco, Estônia, Geórgia, Hungria, Latvia, Lituânia, Polônia, Romênia, Rússia e a Ucrânia, apresenta muitas características semelhantes a Europa/EUA/Canadá. Tem um crescimento populacional baixo (0,24), as expectativas de vida, tanto feminina (76 anos) como masculina (68 anos), estão acima das mundiais, Quase que a totalidade de seus habitantes são alfabetizados (97,47%). Quanto aos

índices de saúde a mortalidade infantil é bem acima do que a da Europa, 16,89% e o número de aids (1,26) está bem abaixo do índice mundial.

Tabela 7: Estatísticas Descritivas da Europa Oriental (n = 14)

Variável	Média	Desvio Padrão
População (mil)	23.300,00	39.109,67
Densidade Populacional (hab/km)	76,70	34,79
População Urbana(%)	62,23	10,58
Expectativa de vida feminina (anos)	76,00	1,11
Expectativa de vida masculina (anos)	67,71	2,33
Alfabetização (%)	97,00	3,76
Crescimento Populacional (%)	0,24	0,32
Mortalidade Infantil (%)	16,89	5,48
Produto Interno Bruto	5.159,79	1.708,70
Calorias	3.477,00	278,92
Aids (nº de casos)	261,92	746,21
Natalidade (nat/1.000)	13,43	1,28
Mortalidade (mort/1.000)	10,82	1,73
Índice de Aids (Aids/10.000)	1,26	3,16
Taxa de natalidade por mortalidade	1,29	0,35
Fertilidade	1,89	0,13
Alfabetização Masculino (%)	99,67	0,50
Alfabetização Feminino (%)	99,33	1,00

O Oriente Médio (tabela 8), por ter seu modo de vida mais reservado, apresenta um índice de aids (0,92%) bastante baixo, porém 41,39% de suas crianças morrem antes de completar um ano. Sua população alfabetizada é de 74,75%. A expectativa de vida para as mulheres é de 71 anos e 67 anos para os homens. A discrepância entre homens e mulheres é revelada quando se verifica o índice de alfabetização por gênero e observa-

se um índice avantajado para os homens. Nesta região a taxa de natalidade é alta, porém a de mortalidade é baixa provocando, assim, uma explosão demográfica.

Tabela 8: Estatísticas Descritivas do Oriente Médio (n= 17)

Variável	Média	Desvio Padrão
População (mil)	17.640,06	22.509,39
Densidade Populacional (hab/km)	126,66	200,57
População Urbana(%)	65,94	21,61
Expectativa de vida feminina (anos)	71,59	4,50
Expectativa de vida masculina (anos)	67,35	3,92
Alfabetização (%)	74,75	15,87
Crescimento Populacional (%)	2,95	1,12
Mortalidade Infantil (%)	41,39	19,18
Produto Interno Bruto	4.957,41	4.057,45
Calorias	3.083,38	253,92
Aids (nº de casos)	55,38	70,79
Natalidade (nat/1.000)	32,71	8,27
Mortalidade (mort/ 1.000)	6,06	1,82
Índice de Aids (Aids/100. 000)	0,92	1,38
Taxa de natalidade por mortalidade	6,06	2,80
Fertilidade	4,72	1,54
Alfabetização Masculino (%)	80,44	14,43
Alfabetização Feminino (%)	66,44	21,39

Ao se analisar as regiões separadamente tem-se uma idéia do porquê dos dados mundiais, pois é um resumo de todas regiões. Através dos resultados obtidos pode-se verificar quais indicadores devem ser melhorados para que se alcance uma qualidade de vida melhor para a população mundial.

5.1 Correlação

A fim de se verificar quais as variáveis são correlacionadas foi feita uma matriz de correlação, que encontra-se no ANEXo I. Com base nesta tabela verificou-se a correlação entre todas as variáveis do banco de dados, contemplado com as informações mundiais. Verifica-se que a densidade populacional está diretamente correlacionada com a população urbana e o número de casos de aids, a população urbana está diretamente correlacionada com a expectativa de vida, tanto masculina como feminina e inversamente correlacionada com a mortalidade. Verifica-se também que quanto maior o índice de alfabetização menor o crescimento populacional. A mortalidade infantil está diretamente correlacionada com a taxa de natalidade e de mortalidade e quanto maior o PIB da região, menor a mortalidade infantil. (Anexo 1)

Para finalizar pode-se averiguar na matriz de correlação que as variáveis são bastante correlacionadas, facilitando assim o uso da imputação dos dados pelo algoritmo EM e regressão linear.

Tabela 9: Resumo da Matriz de Correlação

VARIÁVEL	CORRELAÇÃO POSITIVA	CORRELAÇÃO NEGATIVA
População Urbana	Expectativa de vida (0,743) Alfabetização (0,650) PIB (0,605) Calorias (0,692)	Mortalidade infantil(-0,718) Natalidade (-0,629) Mortalidade (-0,483) Fertilidade (-0,619)
Expectativa de Vida	Alfabetização (0,777) PIB (0,642) Calorias (0,775)	Crescimento populacional (-0,579) Mortalidade infantil (-0,962) Natalidade (-0,862) Mortalidade (-0,696) Fertilidade (-0,836).
Alfabetização	Calorias (0,682) PIB (0,552)	Natalidade (-0,869) Mortalidade Infantil (-0,900) Mortalidade (-0,486) Fertilidade (-0,866) Crescimento Populacional(-0,699)
Crescimento Populacional	Mortalidade infantil (0,602) Natalidade (0,861) Fertilidade (0,840)	PIB (-0,630) Calorias (-0,609)
Mortalidade Infantil	Natalidade (0,865) Mortalidade (0,630) Fertilidade (0,833)	PIB (-0,640), Calorias (-0,777) Alfabetização (-0,900)
PIB	Calorias (0,751) Alfabetização (0,552)	Natalidade(-0,651) Fertilidade(-0,583)
Calorias	Alfabetização (0,682) PIB (0,751)	Natalidade(-0,762) Fertilidade(-0,696)

6. EXEMPLO

Para ilustrar o trabalho será feito um exemplo didático, usando a rotina "*Missing Value Analysis*" do SPSS, e o banco de dados World95. A variável com dados ausentes será a mortalidade infantil, e os países sem esses valores serão a Bolívia e Cuba. As variáveis que serão utilizadas como variáveis independentes são: expectativa de vida feminina, PIB, alfabetização, natalidade, mortalidade e fertilidade, sendo que as três primeiras estão fortemente correlacionadas negativamente com a mortalidade infantil e as demais estão fortemente correlacionadas positivamente.

A tabela 4 servirá de referência para a comparação das estimativas da média e do desvio padrão da mortalidade infantil obtidas com os bancos de dados completados pelas imputações.

Primeiramente, serão ilustrados, através das figuras a seguir, os passos do SPSS para a imputação através do Algoritmo EM:

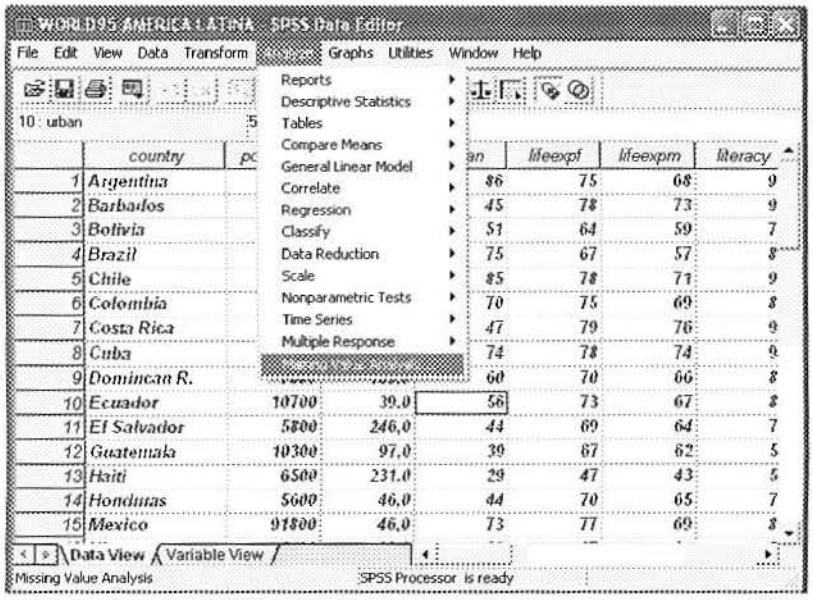

Figura 1: Análise dos Dados

Neste primeiro passo é definida qual a análise que será usada, ou seja, a opção *Missing Value Analysis*.

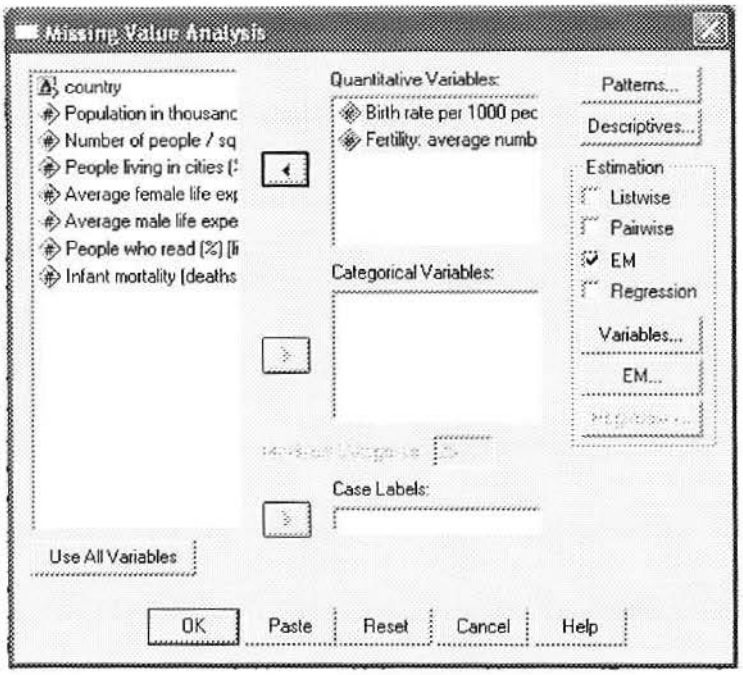

Figura 2: ***Missing Value Analysis***

No segundo passo definem-se quais são as variáveis quantitativas e quais são as qualitativas, definindo ainda qual o método que será utilizado para a imputação.

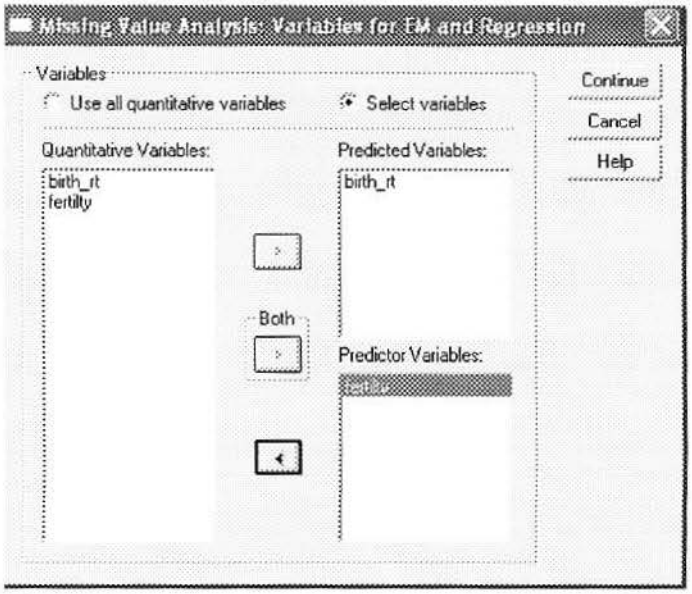

Figura 3: ***Missing Value Analysis*** – Variáveis para EM e Regressão

No passo 3 (figura 3), definem-se quais são as variáveis independentes e qual a variável dependente, dentre as que já foram selecionadas no passo 2..

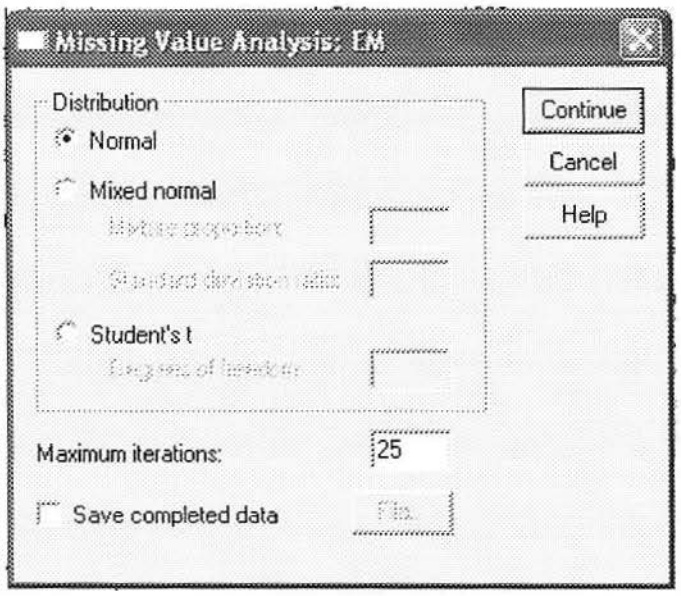

Figura 4: *Missing Value Analysis*: Algoritmo EM

Para completar o processo, seleciona-se qual a distribuição que será utilizada, o número máximo de interações e onde será salvo o novo banco de dados já com os dados imputados.

6.1. Imputação por Algoritmo EM

Neste instante, serão apresentados os resultados da imputação obtidos pelo algoritmo EM, utilizando distribuição normal. Para a escolha das variáveis auxiliares que serão usadas na imputação, verificam-se quais as variáveis mais correlacionadas com mortalidade infantil, conforme tabela abaixo. A tabela completa das correlações da América Latina encontra-se no anexo 2.

Tabela 10: Variáveis correlacionadas com mortalidade infantil

Correlação Positiva		Correlação Negativa	
Fertilidade	0,714	Expectativa de Vida	-0,965
Mortalidade	0,696	Alfabetização	-0,718
Natalidade	0,662	PIB	-0,684

A seguir estão os resultados dos dados imputados, utilizando o Algoritmo EM, com 25 interações (*default* do SPSS) para cada análise feita.

Notação: Nas tabelas que seguem teremos a seguinte notação:

EXP: Expectativa de Vida **NAT:** Natalidade

ALF: Alfabetização **MORT:** Mortalidade

PIB: Produto Interno Bruto **FERT:** Fertilidade

Tabela 11: Valores imputados (em negrito) pelo algoritmo EM para mortalidade infantil (Correlação Negativa)

	EXP	ALF	PIB	EXP x ALF	EXP x PIB	ALF x PIB	EXP x ALF x PIB
Bolívia	**62,8**	**44,5**	**49,7**	**62,4**	**62,5**	**44,8**	**61,0**
Cuba	**19,9**	**23,1**	**44,5**	**19,8**	**19,0**	**23,5**	**21,2**
Estimativas							
Média	38,99	38,28	39,54	38,97	38,94	38,31	38,87
Desvio Padrão	23,35	23,00	23,29	23,33	23,37	22,99	23,35
MCAR (valor-p)	0,000	0,000	0,000	0,000	0,000	0,000	0,000

De acordo com a tabela 11 pode-se ver que em todas as situações o teste de hipóteses, cuja hipótese nula é de que os dados ausentes tem mecanismo de não-resposta MCAR, rejeitaram a hipótese ($p<0,001$). Portanto, assume-se que os dados são MAR.

Com esta verificação, sabe-se de antemão que o resultado da imputação pelo método EM tem um viés nas estimativas (HIPPEL, 2004). Para o conjunto de países da América Latina, o desvio padrão da mortalidade infantil é 24,52 (tabela 4). Portanto, percebe-se que todas as imputações fizeram com que os desvios estimados ficassem abaixo do verdadeiro.

Nota-se ainda que quando é considerada a variável expectativa de vida como variável auxiliar, sozinha ou conjuntamente, as estimativas com os dados imputados são levemente mais próximos dos valores verdadeiros (média=39,11; DP=24,52).Talvez isso aconteça pelo fato da expectativa de vida ter a maior correlação com mortalidade infantil.

Tabela 12: Valores imputados (em negrito) pelo algoritmo EM para mortalidade infantil (Correlação Positiva)

	FERT	MORT.	*NAT*	FERT X NAT	FERT. X MORT	NAT. X MORT	FERT. X NAT. X MORT
Bolívia	**51,6**	**47,7**	**53,7**	**49,0**	**53,9**	**57,9**	**61,0**
Cuba	**16,4**	**37,4**	**17,2**	**18,2**	**22,9**	**21,0**	**21,2**
Estimativas							
Média	38,30	39,11	38,43	38,26	38,72	38,82	38,97
Desvio	23,62	22,92	23,56	23,46	23,16	23,33	23,44
MCAR (valor p)	0,000	0,001	0,000	0,000	0,00	0,00	0,000

A mesma suposição do mecanismo do dados ausentes foi verificada e observa-se que novamente os dados ausentes foram considerados MAR. Pela tabela 12 vê-se que os valores das médias e desvios padrão novamente foram subestimados.

6.2. Imputação por Regressão Linear

A seguir estão apresentados os passos para preencher os dados ausentes utilizando a rotina *Missing Value Analysis* através do método de regressão.

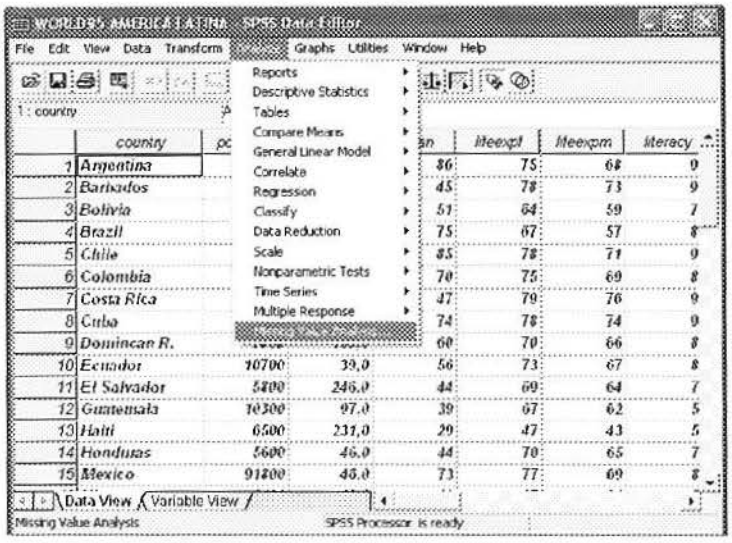

Figura 5: Análise dos dados

O primeiro passo (figura 5) é decidir qual a análise que será feita pelo usuário.

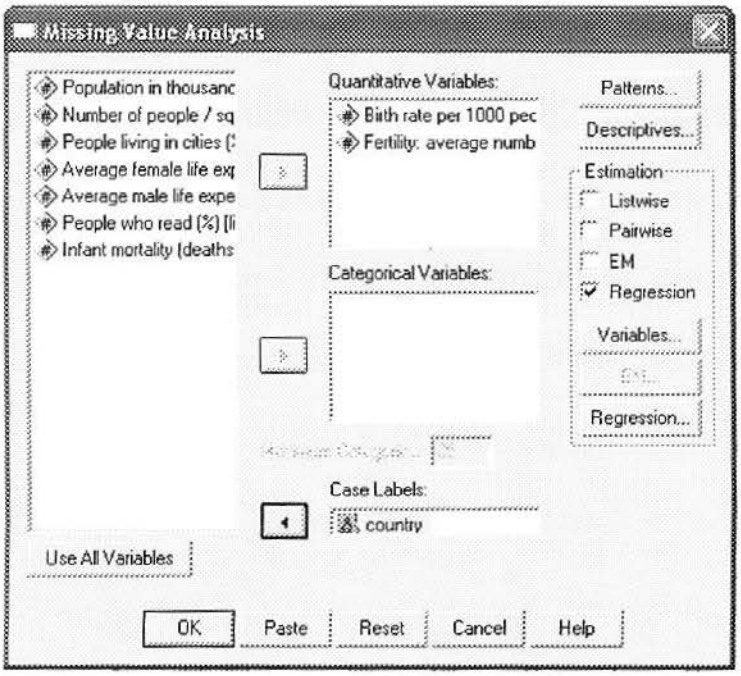

Figura 6: Definição das Variáveis

Na figura 6 (passo 2) define-se as variáveis que serão utilizadas no processo de imputação de dados, classificando-as como quantitativas e qualitativas, e opta-se pelo método de imputação que será utilizado.

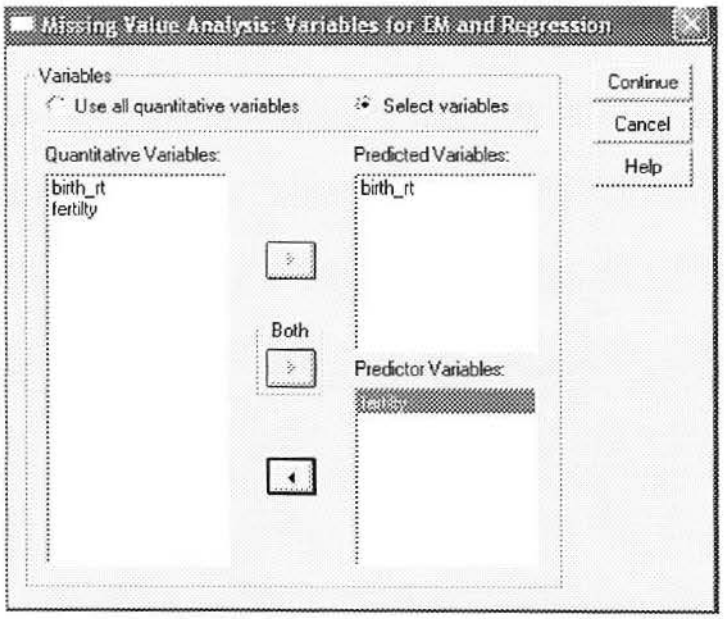

Figura 7: Definição das variáveis (Dependentes e Independentes)

Neste momento define-se qual variável é preditora (independente) e qual é a variável predita (dependente). (Figura 7)

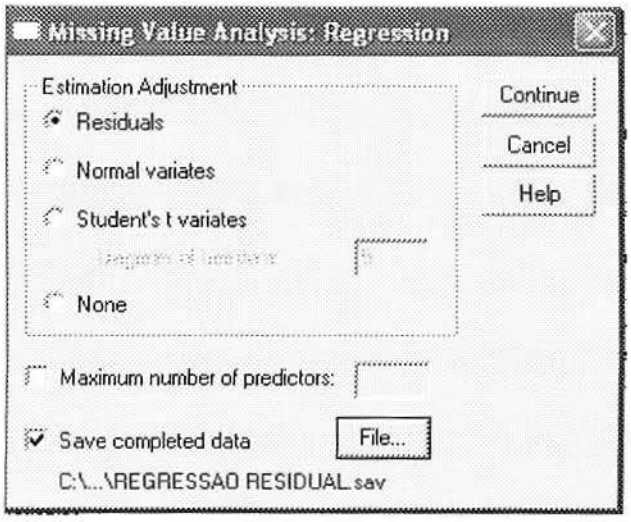

Figura 8: Definição do Ajuste

Para finalizar o processo de imputação utilizando a Regressão Linear define-se o ajuste da estimação e o arquivo onde serão armazenados os dados imputados. (Figura 8)

Na tabela 13 estão os resultados da imputação, obtidos pela Regressão Linear, utilizando a estimação do ajuste residual.

Tabela 13: Valores imputados (em negrito) pela regressão linear para mortalidade infantil (Correlação Negativa)

	EXP	ALF	PIB	EXP X ALF	EXP X PIB	ALF X PIB	EXPEC X ALF X PIB
Bolívia	**75,2**	**47,8**	**79,7**	**63,6**	**59,2**	**47,8**	**61,8**
Cuba	**16,3**	**27,4**	**67,2**	**20,3**	**23,5**	**13,2**	**20,3**
Estimativas							
Média	39,42	38,64	42,05	39,05	38,99	37,96	38,97
Desvio	24,21	22,49	24,66	23,30	22,97	23,05	23,20

De acordo com a tabela 13, verifica-se que as variáveis expectativa de vida e PIB apresentam desvios padrão próximos do valor verdadeiro (24,52). Entretanto, a média da mortalidade infantil tendo o PIB como variável auxiliar ficou acima do valor verdadeiro (39,11), enquanto a estimativa da média quando considerada expectativa de vida ficou bastante próxima da verdadeira média.

Tabela 14:Valores imputados (em negrito) pela regressão linear para mortalidade infantil (Correlação Positiva)

	FERT	MORT	NAT	FERT X NAT	FERT. X MORT	NAT. X MORT	FERT. X NAT. X MORT
Bolívia	**70,3**	**52,3**	**43,0**	**48,9**	**52,3**	**39,2**	**39,2**
Cuba	**6,8**	**52,3**	**9,0**	**9,0**	**33,6**	**41,0**	**54,0**
Estimativas							
Média	38,08	40,04	37,53	37,81	39,15	38,88	39,50
Desvio	25,45	22,63	23,21	23,31	22,48	22,26	22,49

Conforme a tabela 14, vê-se que a estimativa do desvio apresentada pela fertilidade é levemente superior ao desvio verdadeiro. Quando são consideradas as variáveis fertilidade e mortalidade simultaneamente, a média estimada foi muito próxima da verdadeira (39,11).

Analisando-se os resultados das estimativas obtidas pelo método de regressão múltipla utilizando o *Missing Value Analysis*, pode-se verificar que os dados imputados são diversos e são bastante diferentes dos valores reais da mortalidade infantil, Bolívia (75,0) e Cuba (10,2). Porém as estimativas da média e do desvio não foram muito afetadas.

6. 3. Regressão Linear – Manualmente

Para realizar uma regressão linear é necessário seguir os passos que são apresentados nas figuras de 9 a 14.

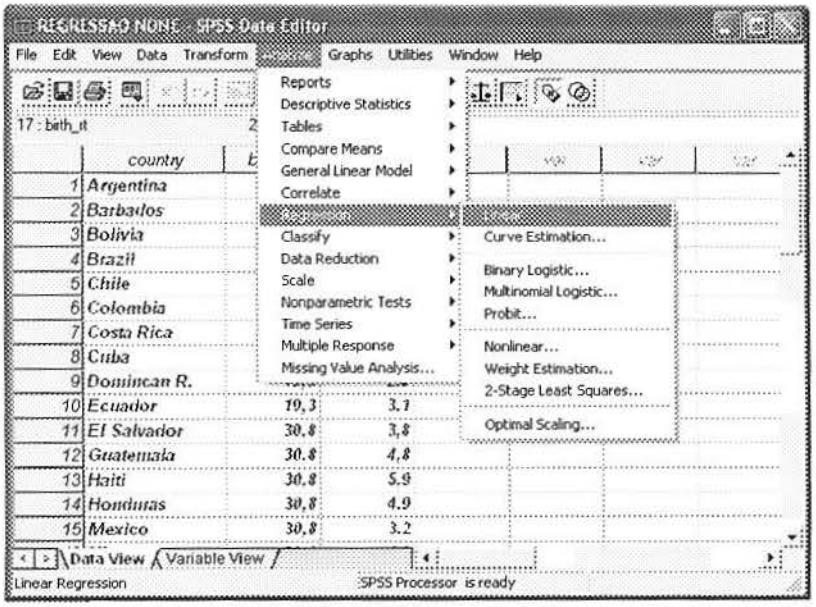

Figura 9: Análise dos dados

O primeiro passo é escolher qual a análise que será feita. Na figura 9 acima vê-se que a opção escolhida deve ser a *regression* e a seguir *linear*.

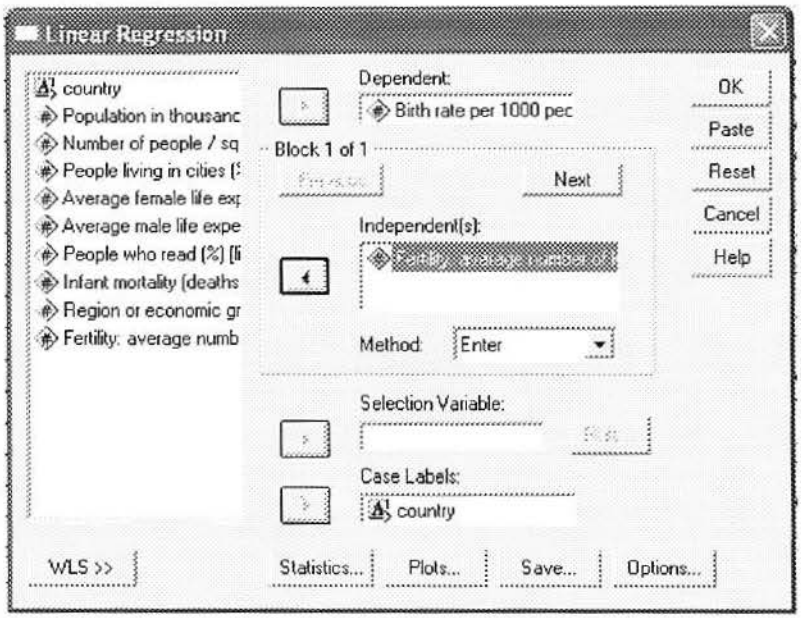

Figura 10: Regressão Linear

No segundo passo, (figura 10), definem-se as variáveis dependentes e as independentes, suas estatísticas e seus gráficos.

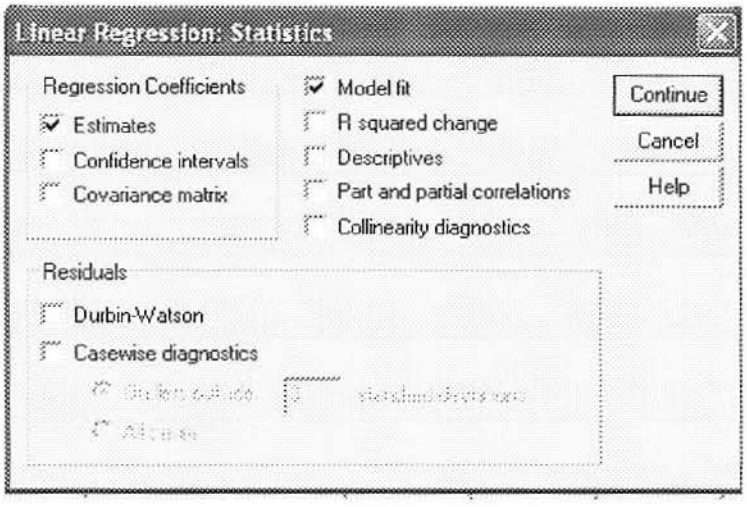

Figura 11: Regressão Linear – Estatísticas.

As estatísticas que podem ser selecionadas são: os coeficientes de regressão e os resíduos. (Figura 11)

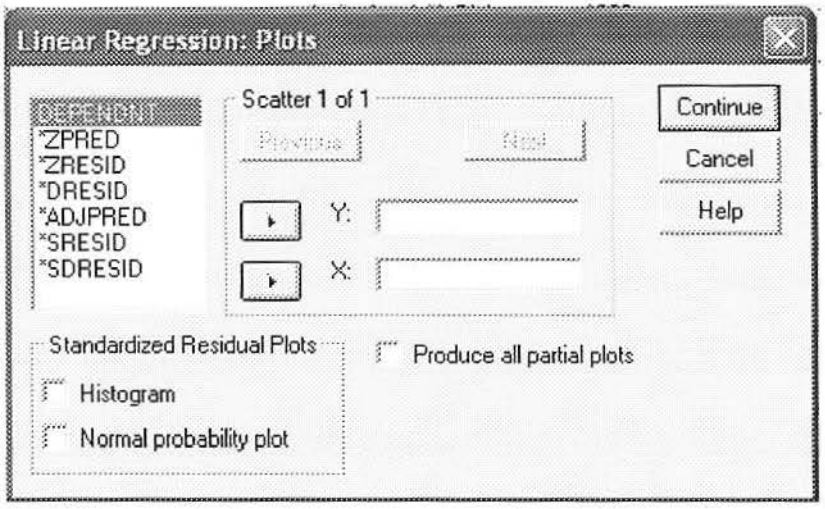

Figura 12: Regressão Linear - Gráficos

Conforme Figura 12, são definidos os gráficos dos resíduos para uma análise posterior.

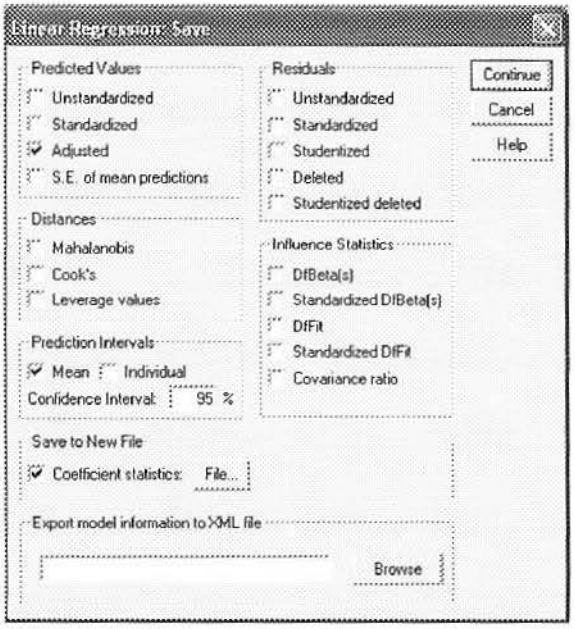

Figura 13: Regressão Linear: Salvar

O SPSS permite o armazenamento de vários valores, por exemplo os valores preditos, os resíduos e outras medidas de diagnóstico. Estes valores serão armazenados no próprio banco de dados com nomes definidos pelo próprio SPSS. (Figura 13)

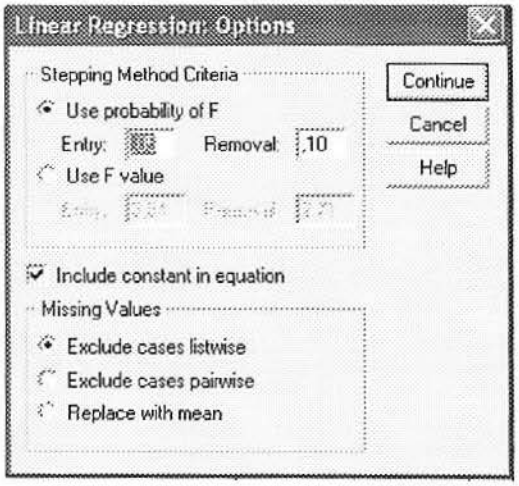

Figura 14: Regressão Linear

A figura 14 mostra o critério de inclusão e exclusão de variáveis no modelo múltiplo. Por *default* o nível de significância para inclusão é de 0,05 e para exclusão é de 0,10.

Tabela 15: Valores imputados (em negrito) pela regressão linear manual para mortalidade infantil (Correlação Negativa)

	EXP	ALF	PIB	EXP X ALF	EXP X PIB	ALF X PIB	EXP X ALF X PIB
Bolívia	**75,93**	**76,44**	**95,28**	**76,3**	**76,6**	**78,6**	**76,1**
Cuba	**32,92**	**55,56**	**89,39**	**33,3**	**29,9**	**58,5**	**32,9**
Estimativas							
Média	40,24	41,34	43,85	40,27	40,13	41,58	40,25
Desvio	23,74	23,94	27,49	23,76	23,84	24,19	23,75

Através dos resultados da imputação, utilizando o método de regressão linear manual, conforme Tabela 15, pode-se ver que a estimativa do desvio, tendo como

variáveis independentes, alfabetização e PIB foi a mais próxima do desvio real (24,52). Pode-se verificar que quando utiliza-se a expectativa de vida como variável independente, as regressões lineares, simples ou múltipla, estimaram desvios padrões próximos dos parâmetros.

Tabela 16: Valores imputados (em negrito) pela regressão linear manual para mortalidade infantil (correlação Positiva)

	FERT	MORT.	NAT	FERT X NAT	FERT. X MORT	NAT. X MORT.	FERT. X NAT. X MORTL.
Bolívia	**90,76**	**84,1**	**96,6**	**89,73**	**84,7**	**87,01**	**91,4**
Cuba	**56,8**	**73,5**	**61,4**	**68,1**	**54,9**	**51,0**	**50,8**
Estimativas							
Média	42,08	42,56	42,58	42,57	41,7	41,63	41,82
Desvio	25,20	25,39	25,93	25,54	24,58	24,71	25,12

A imputação através da regressão linear manual, considerando as variáveis correlacionadas positivamente, verifica-se na Tabela 16, estimou um desvio padrão próximo do parâmetro (24,52) quando consideradas como variáveis independentes a fertilidade e a mortalidade (24,58).

7. CONCLUSÃO

Quando existem valores ausentes no banco de dados, o pesquisador pode optar pela utilização de métodos de imputação já implantados em *softwares*, método "manual" de imputação ou até mesmo desconsiderar os dados ausentes, porém é preciso que se tenha um conhecimento grande do banco de dados a fim de se decidir qual o caminho a ser seguido. Decidir por fazer ou não a imputação dos dados pode ser tão complicado quanto decidir sobre qual a melhor técnica estatística que se aplica aos dados, pois a utilização de um método que não condiga com a realidade dos dados, pode levar a uma conclusão errônea. Este foi o principal motivo para se estudar a utilização da imputação de dados e quais são suas vantagens e desvantagens.

A partir desse trabalho sobre imputação de dados, pode-se concluir que o preenchimento dos dados ausentes, através de um método de imputação, pode trazer ótimos benefícios, pois resulta na possibilidade de fazer-se uma análise com todas as informações, como se o banco estivesse completo.

Para a utilização do banco de dados World95 existente no SPSS foi necessário um conhecimento maior a respeito dos indicadores demográficos mundiais, a fim de se decidir qual o melhor caminho a ser seguido. Foi evidente que ao se trabalhar com a América Latina estaria se estudando a realidade vivenciada nesta região, incluindo o Brasil.

Quando se utiliza método de imputação por regressão linear, tanto pelo *Missing Value Analysis* como pela "regressão manual", verifica-se que as imputações fazem com

que suas estimativas sejam próximas dos parâmetros, evidenciado pelas variáveis mais correlacionadas com mortalidade infantil, que apresentam uma proximidade maior ainda com o parâmetro.

Com os resultados obtidos, após a imputação dos dados, pode-se verificar que a imputação usando a rotina *Missing Value Analysis*, apresentou um melhor resultado para a regressão linear quando comparada com o algoritmo EM. Este resultado pode ter uma relação direta com o mecanismo dos dados perdidos, ou seja, foi verificado que no exemplo os dados perdidos não eram MCAR, gerando a possibilidade de que eles fossem MAR. Esta afirmação foi verificada através do teste de hipótese de Little.

O presente estudo forneceu informações suficientes para que o leitor possa realizar uma imputação usando a rotina *Missing Value Analysis* do SPSS e também mostrou evidência de que a rotina de imputação de dados implementada no SPSS cumpre plenamente sua função, resolvendo de uma forma rápida e clara o processo de preenchimento dos dados ausentes. Entretanto, deve-se ressaltar que o pesquisador antes de decidir pelo método de imputação, deve primeiramente definir qual o mecanismo dos dados perdidos, a fim de não haver um viés nos resultados, pois estes prejudicariam a análise final dos dados.

8. CONSIDERAÇÕES

Assim como se estudam as técnicas de imputação, devem-se estudar as rotinas computacionais criadas para este fim, pois a imputação de dados tornou-se uma técnica muito utilizada e há muitos programas que contêm rotinas que preenchem os dados ausentes. Pacotes estatísticos consideram em suas rotinas de imputação as suposições a serem respeitadas pelas técnicas e por isto o usuário deve conhecer a rotina e os procedimentos que o software está realizando, pois o fácil acesso a esses pacotes faz com que muitos usuários os utilizem sem saber o que o programa está realmente levando em consideração para atribuir este valor.

Indiferente do processo de imputação aplicado para preencher os dados ausentes, não se pode esquecer que a finalidade da imputação é a melhora na análise final. Então todo o cuidado é pouco quando se verifica se as estimativas foram subestimadas ou superestimadas, por que não se pode esquecer que a análise dos dados será feita com base neste novo banco de dados "completo" e não se deseja resultados errôneos.

Convém também salientar que todo pesquisador deve ter extremo cuidado com a formação do banco de dados, por que é com estes dados que, por exemplo, será testada uma nova fórmula de medicação, que uma decisão importante será tomada em políticas públicas, que será implementado um novo modelo de educação infantil... Ou seja, com esta análise pode-se melhorar a qualidade de vida de muitas pessoas e um banco de dados que trará um viés em seus resultados pode trazer muitos prejuízos a esta sociedade. Deve-se ter cuidado quando houver muitos dados ausentes, pois seu resultado de imputação não será de uma "boa" imputação. E ainda, deve-se ter cuidado

especial na formulação do questionário, ao treinamento dos pesquisadores de campo e a realização da pesquisa como um todo, pois só desta maneira tem-se uma redução de dados ausentes nos bancos de dados.

9. BIBLIOGRAFIA

1. Albieri, Sônia. **Ausência de resposta em pesquisa: uma aplicação de métodos de imputação**. Conselho Nacional de Desenvolvimento Cientifico e Tecnológico. Instituto de Matemática Pura e Aplicada. Informes de Matemática. Série D-048/92. Rio de janeiro, 1992.

2. Allison, Paul D. **Mutiple Imputation for Missing Data**. University of Pennsylvania, 2004.

3. Barros Filho, Cyro Rezende. **História Econômica Geral**. Pág 87-129

4. Buntime, W.. **Operations for learning with graphical models**. Journal of Artificial. Intelligence Research, 2:159–225, 1994

5. Chaunnu, Pierre – **Economia da América Latina**, São Paulo, Berttrama Brasil, S/D

6. Dorneles, César Eduardo da Silva. "**Tratamento de missing data por imputação**". Porto Alegre, 1999 – Monografia.

7. Engels, Jean Mundahl & Diehr Paula. (2002). " **Imputation of missing longitudinal data: a comparison of methods**. Journal of clinical Epidemiology 56(2003) 968-976.

8. Fay, Robert. E. (1996). "**Alternative Paradigms for the Analysis of imputerd survey data**". Journal of the American Statistical Association, 91, 490-498

9. Garcia, Carlos Henrique. **Aplicação de equações de modelos Mistos em testes clonais de Eucalyptus SPP**. USP, 2004. Piracicaba, Dissertação de Mestrado.

10. Horton, Nicholas J & Lipsitz, Stuart. (2001). **Multiple Imputation in Practice: Comparison of Software Packages for Regression Models with Missing Variables.** The American Statistician. 244-254.

11. Hippel, Paul T. Von. **Biases in SPSS 12.0 Missing Value Analysis.** The American Statistician, May 2004, vol. 58. Número 2.

12. Koerich. Alessandro L. **Aprendizagem de Máquinas. Redes Bayesianas**. Mestrrado em Informática Aplicada. Pontificia Universidade Católica do Paraná. 2004.

13. Luna, José Eduardo Ochoa. **Algoritmo EM para aprendizagem em redes Bayesiana a partir de dados incompletos**. Departamento de Computação e Estatística. Universidade Federal do Mato Grosso do Sul. Mato Grosso do Sul, Julho, 2004.

14. Moreira Morvan. **Envelhecimento e desenvolvimento humano: as transformações demográficas anunciadas na américa latina**. Modificado em 30/08/05

15. Mattson, Nanna. **An Analysis of a multiple Regresion Imputation Model for the variable hours worked in hek 2002**. Uppsala Universitet Master Thesis. Department of Statistics. 2005

16. Mitchell, T. **Machine Learning**. McGraw Hill, 1997.

17. Mood, F. Graybill, e Boes. D. **Introduction to the Theory of Statistics**. New York, McGraw-Hill, Third Edition, 1974.

18. Raghunathan, Trivellore E. (2004). "**What do we do with missing data? Some options for analysis of incomplete data**"Annu. Rev. Public Health 2004.25: 99-117.

19. Rao, J. N. K. "**On variance estimation imputed survey data**". Journal of the American Statistical Association. June, 1996, Vol 91, Nº 434. Application and case studies.

20. Rubin, Ronald B. (1996). "**Multiple Imputation After 18+ Years**", Journal of the American Statistical Association, 91(nº 434) 473-489.

21. Siegel, Sidney. **Estatística não paramétrica**. 1956. McGraw-Hill

22. Silva, Pedro Luis do Nascimento: "**Crítica e imputação de dados quantitativos utilizando o SAS**". Informes de Matemática. Conselho de Desenvolvimento Cientifico e Tecnológico. Rio de Janeiro. 1989.

23. Wazermann, Claudia (Org). **América Latina cinco Século**. Editora da UFRGS, 2003. pág. 39-77

Sites Pesquisados

1. http://www.bertolo.pro.br/AdminFin/StatFile/Manual_Estatistica.htm (25/06/05)
2. jfsoares@est.ufmg.br (30/08/05)

10. ANEXOS

Anexo 1: Matriz de Correlação – Geral

		População (mil)	Densidade Populacional (%)	População Urbana (%)	Expectativa de Vida Feminina	Expectativa de Vida Masculina	Taxa de alfabetismo (%)	Crescimento Populacional (%)	Mortalidade Infantil	Produto Interno Bruto
População (mil)	Correlação de Pearson	1	-,018	-,175	-,071	-,033	-,064	-,050	**,088**	-,87
	N	109	109	108	109	109	107	109	109	109
Densidade Populacional (%)	Correlação de Pearson		1	,223	,128	,151	,031	-,165	-,145	,201
	N		109	108	109	109	107	109	109	109
População Urbana (%)	Correlação de Pearson			1	,743	,730	,650	-,375	-,718	,605
	N			108	108	108	107	108	108	108
Expectativa de vida Feminina	Correlação de Pearson				1	,982	,865	-,579	-962	,642
	N				109	109	107	109	109	109
Expectativa de vida Masculina	Correlação de Pearson					1	,809	-,502	-,936	,639
	N					109	107	109	109	109
Taxa de alfabetismo (%)	Correlação de Pearson						1	-,699	-,900	,552
	N						107	107	107	107
Crescimento Populacional (%)	Correlação de Pearson							1	,602	-,521
	N							109	109	109
Mortalidade Infantil	Correlação de Pearson								1	-,640
	N								109	109
Produto Interno Bruto										1
	N									109

* Correlation is significant at the 0.05 level (2-tailed).
** Correlation is significant at the 0.01 level (2-tailed).

Anexo 1: Matriz de Correlação – Geral (Continuação)

		Calorias	Aids	Natalidade	Mortalidade	Aids/10000 hab.	Nat/Mort	Fertilidade	Taxa de alfabetismo Masc.(%)	Taxa de alfabetismo Fem.(%)
População (mil)	Correlação de Pearson	-,047	,138	-,039	-,048	-,054	-,051	-,069	,006	-,046
	N	75	106	109	108	106	108	107	85	85
Densidade Populacional (%)	Correlação de Pearson	,067	-,042	-,153	-,121	-,060	-,061	-,162	,085	,029
	N	75	106	109	108	106	108	107	85	85
População Urbana (%)	Correlação de Pearson	,692	,043	-,629	-,483	-,287	-,032	-,619	,587	,612
	N	74	105	108	107	105	107	106	85	85
Expectativa de vida Feminina	Correlação de Pearson	,775	,019	-,862	-696	-,433	-,087	-,838	,777	,819
	N	75	106	109	108	106	108	107	85	85
Expectativa de vida Masculina	Correlação de Pearson	,765	,011	-,805	-,739	-,447	-,011	-,783	,717	,745
	N	75	106	109	108	106	108	107	85	85
Taxa de alfabetismo (%)	Correlação de Pearson	,682	,062	-,869	-,486	-,160	-,271	-,866	,948	,973
	N	74	104	107	106	104	106	105	85	85
CrescimentoPopulacional (%)	Correlação de Pearson	-,609	-,056	,861	-,041	,083	,800	,840	-,619	-,638
	N	75	106	109	108	106	108	107	85	85
Mortalidade Infantil		-,777	-,044	,865	,630	,285	,118	,833	-,809	-,843
		75	106	109	109	106	108	107	85	85
Produto Interno Bruto	Correlação de Pearson	,751	,258	-,651	-,165	-,031	-,338	-,583	,417	,429
	N	75	106	109	108	106	108	107	85	85

* Correlation is significant at the 0.05 level (2-tailed).
** Correlation is significant at the 0.01 level (2-tailed).

Anexo 1: Matriz de Correlação – Geral (Continuação)

		Calorias	Aids	Natalidade	Mortalidade	Aids/10000 hab.	Nat/Mort	Fertilidade	Taxa de alfabetismo Masc.(%)	Taxa de alfabetismo Fem.(%)
Calorias	Correlação de Pearson	1	,167	-,762	-,352	-,245	-,240	-,696	,576	,548
	N	75	75	75	75	75	75	75	59	59
Aids	Correlação de Pearson		1	-,054	,066	,397	-,106	-,046	,085	,096
	N		106	106	106	106	106	106	84	84
Natalidade	Correlação de Pearson			1	,367	,278	,483	,975	-,794	-,835
	N			109	108	106	108	107	85	85
Mortalidade	Correlação de Pearson				1	,523	-,513	,396	-,486	-,510
	N				106	106	108	107	85	85
Aids/100000 pessoas(%)	Correlação de Pearson					1	-,167	,287	-,156	-,170
	N					106	106	106	84	84
Natalidade/mortalidade (%)	Correlação de Pearson						1	,452	-,153	-,148
	N						108	107	85	85
Fertilidade	Correlação de Pearson							1	-,796	-,639
	N							107	85	85
Alfabetismo Masculino (%)	Correlação de Pearson								1	,964
	N								85	85
Alfabetismo Feminino (%)	Correlação de Pearson									1
	N									85

* Correlation is significant at the 0.05 level (2-tailed).
** Correlation is significant at the 0.01 level (2-tailed).

Anexo 2: Matriz de Correlação – América Latina

		População (mil)	Densidade Populacional (%)	População Urbana (%)	Expectativa de Vida Feminina	Expectativa de Vida Masculina	Taxa de alfabetismo (%)	Crescimento Populacional (%)	Mortalidade Infantil	Produto Interno Bruto
População (mil)	Correlação de Pearson	1	-,231	,361	-,028	-,199	,061	-,174	,180	,160
	N	21	21	21	21	21	21	21	21	21
Densidade Populacional (%)	Correlação de Pearson		1	-,451	-,088	-,032	,000	-,525	,049	,518
	N		21	21	21	21	21	21	21	21
População Urbana (%)	Correlação de Pearson			1	,500	,386	,560	-,342	-,452	,285
	N			21	21	21	21	21	21	21
Expectativa de vida Feminina	Correlação de Pearson				1	,974	,808	-,238	-965	,550
	N				21	21	21	21	21	21
Expectativa de vida Masculina	Correlação de Pearson					1	,792	-,186	-,969	,483
	N					21	21	21	21	21
Taxa de alfabetismo (%)	Correlação de Pearson						1	-,508	-,779	,608
	N						21	21	21	21
Crescimento Populacional (%)	Correlação de Pearson							1	,273	-,682
	N							21	21	21
Mortalidade Infantil	Correlação de Pearson								1	-,497
	N								21	21
Produto Interno Bruto										1
	N									21

* Correlation is significant at the 0.05 level (2-tailed).
** Correlation is significant at the 0.01 level (2-tailed).

Anexo 2: Matriz de Correlação – América Latina (Continuação)

		Calorias	Aids	Natalidade	Mortalidade	Aids/10000 hab.	Nat/Mort	Fertilidade	Taxa de alfabetismo Masc.(%)	Taxa de alfabetismo Fem.(%)
População (mil)	Correlação de Pearson	,439	,962	-,212	,007	-,046	-,163	-,197	,054	,068
	N	21	21	21	21	21	21	21	21	21
Densidade Populacional (%)	Correlação de Pearson	-,389	-,141	-,133	,292	,822	-,337	-,114	-,015	,003
	N	21	21	21	21	21	21	21	21	21
População Urbana (%)	Correlação de Pearson	,557	,228	-,651	-,259	-,371	-,265	-,642	,542	,578
	N	21	21	21	21	21	21	21	21	21
Expectativa de vida Feminina	Correlação de Pearson	,707	-,135	-,684	-,784	-,129	,230	-,758	,756	,833
	N	21	21	21	21	21	21	21	21	21
Expectativa de vida Masculina	Correlação de Pearson	,622	-,298	-,620	-,784	-,119	,294	-,97	,750	,808
	N	21	21	21	21	21	21	21	21	21
Taxa de alfabetismo (%)	Correlação de Pearson	,680	-,027	-,830	-,470	,029	-,087	-,839	,986	991
	N	21	21	21	21	21	21	21	21	21
CrescimentoPopulacional (%)	Correlação de Pearson	-,417	-,192	,791	-,316	-,495	,790	,663	-,480	-518
	N	21	21	21	21	21	21	21	21	21
Mortalidade Infantil		-,674	,270	,662	,696	,125	-,196	,714	-,718	-810
		19	19	19	19	19	19	19	19	21
Produto Interno Bruto	Correlação de Pearson	,834	,124	-,658	-,133	,578	-,300	-,600	,581	,617
	N	21	21	21	21	21	21	21	21	21

* Correlation is significant at the 0.05 level (2-tailed).
** Correlation is significant at the 0.01 level (2-tailed).

Anexo 2: Matriz de Correlação – América Latina (Continuação)

		Calorias	Aids	Natalidade	Mortalidade	Aids/10000 hab.	Nat/Mort	Fertilidade	Taxa de alfabetismo Masc.(%)	Taxa de alfabetismo Fem.(%)
Calorias	Correlação de Pearson	1	,325	-,650	-,407	-,194	,111	-,605	,611	,721
	N	21	21	21	21	21	21	21	21	21
Aids	Correlação de Pearson		1	-,157	,114	,083	-,196	-,110	-,049	-,009
	N		21	21	21	21	21	21	21	21
Natalidade	Correlação de Pearson			1	,237	-,129	,489	,953	-,794	-843
	N			21	21	21	21	21	21	21
Mortalidade	Correlação de Pearson				1	,357	-,667	,424	-,438	-487
	N				21	21	21	21	21	21
Aids/100000 pessoas(%)	Correlação de Pearson					1	-,324	-,026	,005	,046
	N					21	21	21	21	21
Natalidade/mortalidade (%)	Correlação de Pearson						1	,327	-,101	-,075
	N						21	21	21	21
Fertilidade	Correlação de Pearson							1	-,800	-852
	N							21	21	21
Alfabetismo Masculino (%)	Correlação de Pearson								1	,956
	N								21	21
Alfabetismo Feminino (%)	Correlação de Pearson									1
	N									21

* Correlation is significant at the 0.05 level (2-tailed).
** Correlation is significant at the 0.01 level (2-tailed).